since $\log p/K$ is small. So the problem has essentially been reduced to finding the number a that satisfies

$$\inf_{\theta} M(a,\theta) = 1 - \frac{\log p}{K}, \tag{9.18}$$

yielding an estimate $m_p = aK$. Expand M about the point $(\mu, 0)$ in a two-term Taylor series with remainder, and find

$$\begin{aligned} m_p &= K(\mu + \Delta a + O(\delta a^2)) \\ &= K\mu + \sigma\sqrt{2K\log p} + K \cdot O\left(\frac{\sigma^2 \log p}{K}\right) \\ &= K\mu + \sigma\sqrt{2K\log p} + O(\log p). \end{aligned} \tag{9.19}$$

Equivalently, you can expand the Chernoff rate $\ell(a)$ near $a = \mu$ as

$$\ell(a) \approx (a-\mu)^2/2\sigma^2. \qquad \spadesuit$$

Since $K/\log p$ is large, the error term in 9.10 is $O(\log p) < \sqrt{K\log p}$.

Corollary 9.11. *If $F(x)$ is IFR, then the estimate of Exercise 9.10 implies*

$$E(T) = K\mu + \sigma\sqrt{2K\log p} + O(\log p). \tag{9.20}$$

Proof. This is immediate from Exercise 9.10 and Lemma 9.9. ■

Comparing this with the result of Exercise 9.5, we see that in this limit, the Normal law derived from the central limit theorem and the estimate from Chernoff's Theorem agree to the lowest order. (In the jargon of matched asymptotic expansions, the inner expansion of large deviations matches the outer expansion of the central limit theorem). This formal calculation can be made precise: this is done, for example, in [Ne1, Ne2] and [Ree]. The delicate point here is the errors in the approximations, since generally the location where a function takes a value is very sensitive to changes in the function.

There are some circumstances where $K/\log p$ might not be large even though K is large. We can still use the theory of large deviations to estimate $\mathbb{E}(T)$ in any particular case, although there does not seem to be a general statement corresponding to Exercise 9.10.

Exercise 9.12. Show that for the exponential distribution, when $\log p/K$ is large then

$$m_p \approx \frac{\log p}{\lambda} + \frac{K}{\lambda}\log\frac{\log p}{K}. \qquad \spadesuit$$

The analysis of the model with random overhead is subsumed by our previous analysis, in the following way. Suppose that it takes processor i time $v(i)$ to access the job queue. The makespan is then given by (9.9), but with $\tilde{Y}_i \stackrel{\triangle}{=} Y_i + v(i)$ replacing Y_i. Let $\tilde{m}_p$ be the characteristic maximum of $\tilde{Y}_i$.

[BLS] it is shown that if the initial rollback is large, then (asymptotically) all rollbacks initialized by this rollback will be no larger.

Let us now apply this theorem to get concrete information about the stability of a simple system. Suppose we can model the changes in rollback size as random variables that take the values 1 and $-K$ with probability q and $(1-q)$ respectively. The mean change in rollback size is $q - K(1-q)$, which is negative whenever $q < \frac{K}{K+1}$.

Example 10.3. For $K = 1$, we have

$$\ell(0) \stackrel{\triangle}{=} \sup_{\theta} \left[-\log(qe^{\theta} + (1-q)e^{-\theta}) \right].$$

Differentiating $qe^{\theta} + (1-q)e^{-\theta}$ to locate the maximum, we get a quadratic equation in e^{θ} and substituting the solution $e^{\theta^*} = \sqrt{1-q}/\sqrt{q}$ yields

$$\ell(0) = -\log\left(2\sqrt{q(1-q)}\right).$$

This implies that for $b = 2$ the critical value of q that separates the stable and unstable regions is

$$q = \frac{2-\sqrt{3}}{4} \approx 0.066987.$$

Example 10.4. For $q = 1/2$,

$$\ell(0) = -\log\left(\frac{1}{2} K^{\frac{1}{K+1}} \left(1 + \frac{1}{K}\right)\right).$$

The stability condition for $b = 2$ is therefore

$$K^{\frac{1}{K+1}} (1 + 1/K) < 1.$$

However, this condition does not hold for any finite value of K. Therefore, it is not possible to obtain stability when $q = 1/2$!

Exercise 10.5. Verify the calculations of Examples 10.3 and 10.4. Compute the critical values of q for $b = 2$, as a function of K. Compare to the critical value obtained from the naïve requirement that the mean step size is negative. ♠

Example 10.6. For $K = 1$, the critical value of q as a function of b is

$$q^* = \frac{1}{2}\left(1 - \sqrt{b^2 - 1}/b\right).$$

The following qualitative statements are immediate:

(i) q^* increases to $1/2$ as b decreases to one.

(ii) q^* increases to $1/2$ much more slowly than b decreases to one.

(iii) For b large, $q^* \approx 1/\left(4b^2\right)$.

It is also possible to obtain some explicit expressions for $\sum_{i=1}^{\infty} \mathbb{E}z(n, x)$; see [BLS].

$M/M/1$ process, which was the subject of §7.3. (This is an $M/M/1$ process but without the boundary, so that $\lambda_2 \equiv \mu$.) In order to avoid excessive repetition, we shall assume that you have mastered the ideas of §7.3.

Here is how the rest of the chapter is organized. In §11.2 we give formal derivations of most of the results of this chapter. This is in keeping with our point of view as expounded in the Introduction to the Applications: derive first, justify later! Most of the remainder of the chapter is concerned with justification (that is, proofs). Sections §11.3 and §11.4 give a derivation of Kurtz's Theorem and a derivation of the large deviations principle for the $M/M/1$ queue, respectively. This is mostly for pedagogic purposes, since Theorems 8.47 and 8.69 establish the principle. However, we don't suppose that everyone who reads this chapter will have gone through all the previous chapters, so we include a fairly simple and straightforward approach. Once we have the large deviations principle established, we proceed to justify the Freidlin-Wentzell theory for the $M/M/1$ process. Then, after justifying using these calculations to calculate the probability of hitting isolated points, we finish with a new calculation: how long are the busy periods of an $M/M/1$ queue, and how large can the process get during long busy periods?

11.2. Heuristic Calculations

We now follow the steps for analyzing a system using large deviations as given in the Introduction to the Applications. We already have a Markov model of the system at hand, and have done the appropriate scaling. The next step is to calculate the system's likely behavior by calculating $z_\infty(t)$. We have by (5.7) or more correctly by Exercise 8.13,

$$\frac{d}{dt} z_\infty = \begin{cases} \lambda - \mu & \text{if } z_\infty > 0 \\ 0 & \text{if } z_\infty = 0 \,. \end{cases} \tag{11.3}$$

The second line in (11.3) comes from (8.11); it can be understood intuitively as a statement that a stable $M/M/1$ queue tends to stay bounded once it becomes empty. Therefore,

$$z_\infty(t) = \max\left(z_\infty(0) - t\,(\mu - \lambda), 0\right), \tag{11.4}$$

or equivalently, with $T^* = z_\infty(0)/(\mu - \lambda)$ and $z_\infty(0) \geq 0$,

$$z_\infty(t) = \begin{cases} z_\infty(0) - (\mu - \lambda)t & \text{if } t \leq T^* \\ 0 & \text{if } t \geq T^* \,. \end{cases} \tag{11.5}$$

That is, the scaled $M/M/1$ process tends to drift to zero, and once there it stays.

The next step is to formulate some large deviations problems. The local rate function ℓ for the $M/M/1$ queue is, by (8.12) or by Theorem 11.13,

$$\ell(x, y) = \begin{cases} \sup_\theta \left\{\theta y - \lambda(e^\theta - 1) - \mu(e^{-\theta} - 1)\right\} & \text{for } x > 0, \text{ or } x = 0 \;\&\; y > 0; \\ 0 & \text{for } x = 0 \text{ and } y = 0; \\ \infty & \text{for } x < 0, \text{ or } x = 0 \text{ and } y < 0. \end{cases} \tag{11.6}$$

(remember that by (C.6), $d\ell/dy = \theta^*$ and use Exercise 7.24). That is, we have found that

$$\mathbb{P}_{ss}\left(z_n \in (a, b)\right) = e^{-nV(a)+o(n)},$$

or put another way,

$$\mathbb{P}_{ss}\left(x(t) \in (na, nb)\right) = \left(\frac{\lambda}{\mu}\right)^{(na+o(n))}.$$

By Exercise 6.91 we can set $b = \infty$ and get

$$\mathbb{P}_{ss}(x(t) \geq na) = \left(\frac{\lambda}{\mu}\right)^{(na+o(n))}.$$

(The astute reader will note that the $o(n)$ term is actually quite small for the $M/M/1$ queue.)

Another immediate consequence of this analysis and of Theorem 6.17 is the following estimate of escape times. Let $\tau(n; a)$ be the first time that $z_n(t) \geq a$.

Corollary 11.1. *For a stable $M/M/1$ queueing process and any $a > 0$ and $\delta > 0$ we have, uniformly over $z_n(0) \in [0, a - \delta)$,*

$$\lim_{n\to\infty} \mathbb{P}_{z_n(0)}\left(\frac{\log \tau(n; a)}{n} \in (V(a) - \delta, V(a) + \delta)\right) = 1.$$

We can also compute transient distributions for $z_n(t)$. For example, what is $\mathbb{P}_x\left(z_n(t) = y\right)$? That is, what is the probability that there are ny customers in the queue after time nt, if the queue starts with nx customers? Clearly we need to calculate

$$\varphi_t(x, y) \stackrel{\triangle}{=} \inf_{F(x,y,t)} I_0^t(r)$$

for $t > 0$ (see Definitions 6.14) where

$$F(x, y, t) = \left\{r \ : \ r(0) = x, \ r(t) = y\right\}.$$

A moment's thought will show that there are two types of paths that go from x to y: paths that avoid the boundary $z = 0$, and paths that hit the boundary. For the first type of path we know from Lemma 5.16 that the cheapest way to go from x to y is in a straight line. However, for the second type of path, there are several possibilities. We can easily see that optimal paths that touch the boundary must be "U" shaped: a straight line segment from x to zero, possibly a segment where the path remains at zero, then a straight line segment from zero to y. This is because the first and last segments have costs identical to the cost of the free process of §7.3, for which all optimal paths are straight line segments. Furthermore, the time spent at zero costs nothing, while the cost of making an excursion away from zero would have positive cost, since $\rho < 1$.

Now consider the lowest cost path from x to y that touches the boundary, where we place no restrictions on the time t the path takes. We can complete the segment

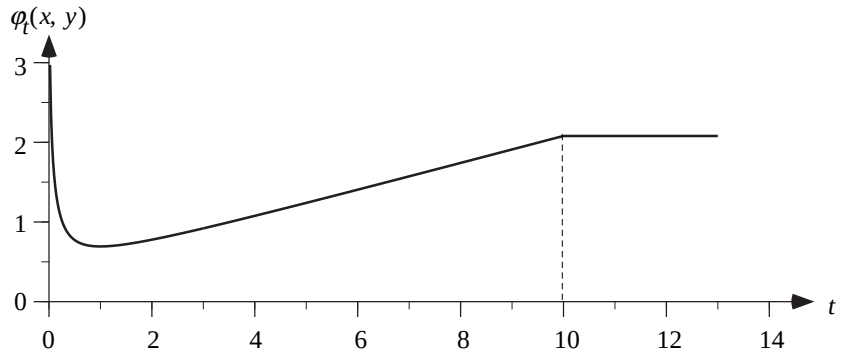

Figure 11.5. $\varphi_t(x, y)$ for the $M/M/1$ queue with $x = 2$, $y = 3$, $\lambda = 1$, $\mu = 2$.

This concludes our heuristic analysis of the $M/M/1$ queue via large deviations. We now begin the justifications and more detailed analysis of the system.

11.3. Most Probable Behavior

In this section we prove that Kurtz's Theorem holds for the $M/M/1$ queue, with $z_\infty(t)$ defined by any of (11.3)–(11.5). If you read Chapter 8 you can, of course, skip this section. We prove the theorem holds by comparing the $M/M/1$ process with the free process. The free process was defined in §7.3. It is the $M/M/1$ queue without a boundary. That is, it is the difference of a Poisson (λ) and an independent Poisson (μ) process. We denote the free process by $\zeta(t)$. The transition structure of $\zeta(t)$ is given by

$$\begin{aligned} e_1 &= +1 \qquad \lambda_1(x) = \lambda \\ e_2 &= -1 \qquad \lambda_2(x) = \mu. \end{aligned} \tag{11.11}$$

It is obvious that this process satisfies the conditions of Kurtz's Theorem and of the large deviations theorems of Chapter 5. At any fixed t, the random variable $\zeta(t)$ is the difference of Poisson (λt) and Poisson (μt) random variables. In §7.3 we showed that the rate function for the free process is

$$\ell(x, y) = \ell(y) = y \log \frac{y + \sqrt{y^2 + 4\lambda\mu}}{2\lambda} + \lambda + \mu - \sqrt{y^2 + 4\lambda\mu}.$$

We make the large deviations scaling

$$\zeta_n(t) \stackrel{\triangle}{=} \frac{1}{n}\zeta(nt).$$

[We call the free process $\zeta_n(t)$ so as not to confuse it with the scaled $M/M/1$ process $z_n(t)$.]

Here is an outline of the argument we use:

1. The time it takes for $z_n(t)$ to go from an initial point $x > 0$ to zero is about $x\,(\mu - \lambda)^{-1}$. We obtain the rate function for deviations from this expected time.

Finite M/M/1 queue.

Another important question is the effect of a boundary at a level M for some $M > 0$. That is, in practice a queue has a finite capacity, and we'd like to know how this finiteness affects the queue's statistics. From the large deviations point of view, this inclusion is almost a triviality. We simply put a barrier at $z_n = M$ (equivalently at $x = nM$), for example, by setting

$$\lambda(x) = \begin{cases} \lambda & \text{if } x < nM; \\ 0 & \text{if } x = nM, \end{cases}$$
$$\mu(x) = \begin{cases} \mu & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Then all the calculations we have done do not change at all so long as $\lambda < \mu$. We still have

$$\mathbb{P}_{ss}(z_n \geq a) = e^{-nV(a)+o(n)}, \tag{11.39}$$

where $V(a) = a \log \frac{\mu}{\lambda}$, as long as $a < M$. Furthermore we have

$$\mathbb{P}_x(z_n(t) = y) = e^{-n\varphi_t(x,y)+o(n)}, \tag{11.40}$$

so long as x and y are both smaller than M, where $\varphi_t(x, y)$ is given by Theorem 11.3 for $x, y \in [0, M]$. We can even extend the analysis to the case $\lambda > \mu$ when we have a barrier at M: simply consider the process $y(t) = M - x(t)$, and we see that $y(t)$ is a stable $M/M/1$ queue with a barrier at M.

Exercise 11.27. What happens if $\lambda = \mu$ when there is a barrier at M? (Answer: $V(a) = 0$ for all $a \in [0, M]$, but $\varphi_t(x, y)$ has the same expression as given in Theorem 11.3, but with $t^* = \infty$. This is because $\varphi_t(x, y) \to 0$ as $t \to \infty$.) ♠

11.10. End Notes

The $M/M/1$ queue is called, confusingly enough, an Erlang system in the older literature. There is a voluminous literature on it and its close relatives, for both steady state and transient analysis. For a classical approach to the $M/M/1$ queue, Takàcs [Ta1] may still be the best reference. For a modern point of view see Massey [Mas]. For more than you ever wanted to know about it and its relatives, see Cohen [C1]. Other books that have worthwhile discussions include [Co, Kl], and any book with the word "queueing" in the title probably has something to say on the subject.

Our result on the transient behavior of the $M/M/1$ queue, Theorem 11.3, may be extracted from [AKW], once the appropriate question is asked. Also, our result on long busy periods can be extended and improved using either more advanced large deviations or other methods. For example, the distribution of the queue length (conditioned on being in a busy period that started a long time ago) tends to a limit that is calculable. The fact that $M/M/1$ queues fill by switching their arrival rate and service rate (Theorem 11.18) can be made more precise by

We see that (13.33) is the equation of either a hyperbolic cosine or a hyperbolic sine, depending on whether c is negative or positive. But the boundary conditions $r(0) = r(T)$ eliminate the hyperbolic sine from consideration, so $c < 0$.

Exercise 13.10. Check that (13.35) gives either a hyperbolic sine or a hyperbolic cosine. Prove that when $q < C$, $c < 0$ implies that $K > \frac{\mu}{1-C} - \frac{\lambda}{C} > 0$. Hint: substitute and use $\cosh^2 x - \sinh^2 x = 1$. ♠

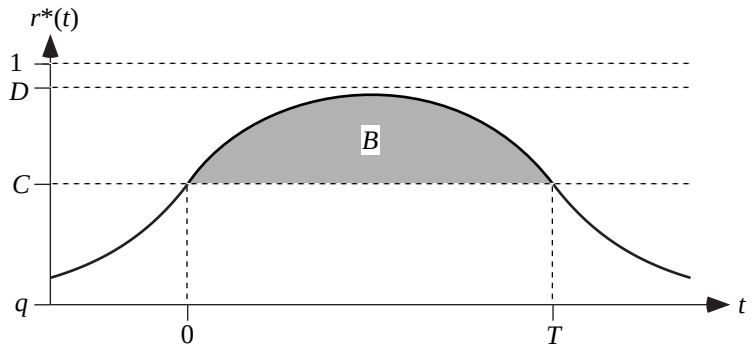

Figure 13.11. $r^*(t)$ for a given value of B. See §13.4 for the meaning of D.

Now denote $r_K(t) = x\cosh(yt + z) + w$. Since cosh is symmetric, the conditions $r(0) = r(T) = C$ imply $y > 0$ and $z < 0$, and we have

$$\begin{aligned} r_K(t) &= x\cosh(yt+z) + w \\ y &= \sqrt{a} \qquad\qquad w = b \\ x &= -\sqrt{-c} \qquad\qquad T = -2z/y \\ \cosh(z) &= (C-w)/x \quad \text{and} \quad z < 0. \end{aligned} \tag{13.37}$$

Given K we can now solve (13.37) for the constants y, w, x, z, T in that order (provided K is such that $c < 0$) to obtain the optimal path r_K.

Our next goal is to find K. We can obtain an explicit expression for

$$\mathbf{B}(r_K)(T) = \int_0^T (r_K(t) - C)\, dt = \frac{2z}{y}(C-w) - \frac{2}{y}\sqrt{(w-C)^2 - x^2}\,. \tag{13.38}$$

Although this provides an explicit expression, it is too messy to be useful (or to be worth writing down . . .). However, we can perform a numerical search for the proper value of K which is determined by $\mathbf{B}(r_K)(T) = B$ (see Figure 13.22). This is particularly simple because of the following properties of $\mathbf{B}(r_K)(T)$ as a function of K, which we establish below:

1. It is monotone decreasing to 0 as $K \to \infty$.
2. It reaches infinity for a fixed positive value of K (see Exercise 13.10 and Theorem 13.16).

We summarize the results we have established concerning the solution of the Lagrange variational problem.

minimal cost curve makes zero area; that is, the curve is a degenerate one, a constant identically equal to C. There are at least three ways of obtaining this result: by straightforward but tedious calculation, by dividing time into intervals where the curve is above or below C and noting that the curvature must be nonnegative in the portion below C but nonpositive in the portion above, or by noting that it is cheaper for the curve to hold at any fixed level below C than any level above C and so we may lower cost by using the convexity of $\ell(x, 0)$ with respect to x. The details are left to the interested reader.

7. Maximum activity in a long busy period.

Just as in the $M/M/1$ queue, we can show that the most likely way that the buffer has a long busy period is for $z_n(t) \approx C$ for the duration of the period. Furthermore, this means that conditioned on a busy period being long, the maximum level of the buffer is $o(n)$, so conditioning on high buffer content is very different than conditioning on a long busy period. Here is a precise statement of the result.

Theorem 13.36. *Given any $\varepsilon > 0$ and $T > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}_{ss}\left(\max_{0\le t\le T} b_n(t) \ge \varepsilon \;\middle|\; \min_{0\le t\le T} b_n(t) > 0 \right) = 0.$$

13.6. Justification

In this section we check that a large deviations principle holds for $z_n(t)$. We also verify that the Freidlin-Wentzell theory applies to the calculation of fluctuations of $b_n(t)$. We prove Lemma 13.15, and finally, we justify the use of a Lagrange multiplier in calculating the minimal cost path that makes a given buffer B.

We begin with the large deviations principle for $z_n(t)$. Our process does not satisfy the hypotheses of Chapters 5 and 8, since there are two boundaries, at zero and at one, and at each of these places the logarithms of the jump rates $\log \lambda_i(x)$ are not bounded: at zero the service rate μx goes to zero, and at one the arrival rate $\lambda(1-x)$ goes to zero. This is exactly the type of singularity encountered in Erlang's model at $x = 0$, and the same argument will prove the large deviations principle in this case as well. We follow the argument of §12.6, subsection "The large deviations principle." The idea is to construct a process $z_n^\varepsilon(t)$ that is identical to $z_n(t)$ whenever z_n is more than ε from a boundary, but satisfies a large deviations principle with a rate function ${}_\varepsilon I_0^T(r)$ that is nearly the same as $I_0^T(r)$. The process we use has the following jump rates and directions:

$$\begin{aligned} e_1 &= 1 \qquad \lambda_1^\varepsilon(x) = \begin{cases} \lambda(1-x) & x < 1-\varepsilon \\ \lambda(1-\varepsilon) & 1-\varepsilon \le x < 1 \\ 0 & x = 1 \end{cases} \\ e_2 &= -1 \qquad \lambda_2^\varepsilon(x) = \begin{cases} \mu x & x \ge \varepsilon \\ \mu\varepsilon & 0 < x \le \varepsilon \\ 0 & x = 0 \end{cases} \end{aligned} \tag{13.67}$$

The rate function for the system is defined by

$$I_0^T(\vec{r}) = \int_0^T \ell(\vec{r}(t), \vec{r}\,'(t))\,dt$$

$$\ell(\vec{x}, \vec{y}) = \sup_{\vec{\theta}\in\mathbb{R}^k} \left(\left\langle \vec{\theta}, \vec{y} \right\rangle - \sum_{i=1}^{k} \left[\lambda_i(f_i - x_i)\left(e^{\theta_i} - 1\right) + \mu_i x_i \left(e^{-\theta_i} - 1\right) \right] \right).$$

It is easy to see that $\ell(\vec{x}, \vec{y}) = \sum_{i=1}^k \ell_i(x_i, y_i)$, where

$$\begin{aligned} \ell_i(x_i, y_i) = y_i \log \frac{y_i + \sqrt{y_i^2 + 4\lambda_i(f_i - x_i)\mu_i x_i}}{2\lambda_i(f_i - x_i)} + \lambda_i(f_i - x_i) \\ + \mu_i x_i - \sqrt{y_i^2 + 4\lambda_i(f_i - x_i)\mu_i x_i}\,. \end{aligned} \tag{13.135}$$

The reason that the ℓ-function splits into a sum of independent pieces is that the process is a sum of independent pieces: each dimension is (statistically) independent of the others. This enables us to piece together the solution of various problems from a number of simple parts.

This split enables us to prove the large deviations principle for $\vec{z}_n(t)$ from the corresponding large deviations principle for the components of $\vec{z}_n(t)$. The lower bound follows directly from the lower bound for the components. Using the simple inequality for $\vec{y} \in \mathbb{R}^k$

$$|\vec{y}| \le \sqrt{k} \max |y_i|$$

we have for any T and $\varepsilon > 0$

$$\begin{aligned} \mathbb{P}_{\vec{x}} \left(\sup_{0\le t\le T} |\vec{z}_n(t) - \vec{r}(t)| < \varepsilon \right) &\ge \prod_i \mathbb{P}_{x_i} \left(\sup_{0\le t\le T} \left| z_{n,i}(t) - r_i(t) \right| < \frac{\varepsilon}{\sqrt{k}} \right) \\ &\ge e^{\left(-n \sum_i \int_0^T \ell_i(r_i, r_i')\,dt + O(n\varepsilon) + o(n) \right)} \\ &= e^{-nI_0^T(\vec{r}) + O(n\varepsilon) + o(n)}. \end{aligned}$$

The upper bound is almost as easy. Recall the definition

$$\Phi_{\vec{x}}(K) = \left\{ \vec{r} : \vec{r}(0) = \vec{x},\ I_0^T(\vec{r}) \le K \right\}.$$

Define the closed set

$$S_{\vec{x}}(\varepsilon, K) \stackrel{\triangle}{=} \left\{ \vec{r}(t) : d\left(\vec{r}, \Phi_{\vec{x}}(K)\right) \ge \varepsilon \right\}.$$

Then define $\vec{y}_n(t)$ as in Chapter 5 as the piecewise linear interpolation of $\vec{z}_n(t)$. Then for each i and $K_i \ge 0$, the large deviations upper bound for the components $z_{n,i}$ of $\vec{z}_n$ gives

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(z_{n,i} \in S_x(\varepsilon, K_i) \right) \le -K_i. \tag{13.136}$$

cost over $\vec{y} \in \{\langle \vec{y}, \vec{a} \rangle = C\}$ gives the asymptotic expression for $I^*(B)$ for B large. This completes our discussion of the solution of the variational problem for the multiple-class model.

We wish to point out that flow control schemes for the multiple class model have exactly the same efficacy as they did in the simple source case. That is, cutoff flow control helps only when it keeps the system from reaching the level $\vec{D}$, and then it has a calculable effect; rate-changing schemes have the same sort of effect on level crossing an buffer sizes, etc.

13.9. End Notes

As the chapter's title implies, the class of models we investigated was proposed and largely solved by Anick, Mitra, and Sondhi in their seminal work [AMS]. The approach presented in this chapter was initiated by us in [We2]. The present exposition largely follows that work, with many corrections and amplifications. The subject has been studied extensively of late by many investigators through a variety of methods. See, for example, [Mo1, EM1, EM2, EMS, KuM, BoD].

We glossed over the point that there are several time scales in a good model of traffic. We examined only the time scale of packet generation and transmission. We explicitly modeled the time scale of bits as fluid, taking a limit that might better be left unevaluated. We also assumed that each source was connected to the system forever, and was always in the same sort of activity mode. A better model would either have sources enter and leave the system, or would have collections of states with extremely long holding times within the collection; some could represent being quiescent, or disconnected, while others could represent various activity levels. David Tse [Tse] addressed this problem using large deviations in his Ph.D. dissertation.

The development of "effective bandwidth" theory has been a major effort in recent years. In some cases it can be viewed as a scaling property of the large buffer asymptotics for the multiple source model. Specifically, there are constants ρ_i such that as $B \to \infty$, $I(B)/B \approx \sum_i f_i \rho_i$ (see (13.131). Effective bandwidths are used in admission control and system design. (For another innovative approach to admission control, see [CKW].) But since we did not contribute to this theory, we did not develop it here. It was initiated in [Hu1] for the zero-buffer case, and by [GH] and [GAN] for the large buffer case, and has also been investigated by [CW, KWC, EM3, KeF, Wh2], and many others.

The models of packet traffic we considered are the beginning of many directions of current research, and the interested student could probably make a contribution just by following some of the directions alluded to above. There are other sorts of models that can be analyzed without undue difficulty. Whenever the underlying source model is reversible then the level crossing problem, at least, can be solved since the most probable path to any given point is the time reversal of $\vec{z}_\infty$ from that point. The small- and large-B asymptotics should then be solvable, if not more.

14.8. End Notes

The analysis of slotted Aloha models in this chapter is based on an AT&T internal report by Albert Greenberg and A. Weiss. This analysis is largely subsumed by the previous work of Cottrell, Fort, and Malgouyres [CFM]. Furthermore, many other people have investigated Aloha via large deviations; see, e.g., Maier [Mai] and his references.

There is a great deal more work that can be done in these multiple-access channels. David Aldous [Al] has shown, based on a large deviations analysis of a ball-and-bucket model, that the popular Ethernet protocol is unstable when subjected to Poisson arrivals. Goodman et al. [GGM] showed that a finite-population Ethernet model is stable for low enough arrival rates. Many adaptive schemes (usually called "controlled Aloha") have been proposed, and most work pretty well. For a good overview of the field, see Rom and Sidi [RS]; also, the IEEE Trans. Info. Theory volume 31 (1985) has many articles on the subject. Even though it is not hard to think of improvements to the Aloha protocol, it survives because of its simplicity and adaptability.

Extremal interior path.

For paths that remain in the interior (i.e., where $x_1 > 0$ and $\vec{x} \notin E$), ℓ is given by (15.12). The strict convexity of ℓ and Lemma 5.16 guarantee that the lowest cost interior path from $\vec{0}$ to the line E must be a straight line. So, suppose our path is of the form

$$\vec{r}(t) = (y_1, y_2) \cdot t \quad \text{for} \quad 0 \le t \le \frac{1}{y_1 + y_2}$$

(so that indeed it ends on E). By (15.14) the minimal cost of a straight line path is

$$\min_{y_1 \ge 0, y_2 \ge 0} \frac{\ell(\vec{y})}{y_1 + y_2} = \min_{\left\{ \begin{smallmatrix} \lambda_i(1), \mu_1(1) \ge 0 \\ \lambda_1(1) \ge \mu_1(1) \end{smallmatrix} \right\}} \frac{h(\lambda_1(1), \lambda_1) + h(\lambda_2(1), \lambda_2) + h(\mu_1(1), \mu_1)}{\lambda_1(1) - \mu_1(1) + \lambda_2(1)}, \tag{15.15}$$

Since $\partial h(a, b)/\partial a \to -\infty$ as $a \to 0$, there must be a minimum where the parameters $\lambda_1(1)$, $\lambda_2(1)$, and $\mu_1(1)$ are strictly positive. Therefore, we are justified in setting partial derivatives of (15.15) with respect to $\lambda_1(1)$, $\lambda_2(1)$, and $\mu_1(1)$ to zero to find the extremal point. From this we obtain

$$\frac{\lambda_1}{\lambda_1(1)} = \frac{\lambda_2}{\lambda_2(1)} = \frac{\mu_1(1)}{\mu_1}. \tag{15.16}$$

Since there is only one point where the derivatives vanish, it must be the global minimum.

We could have derived the relation $\lambda_1/\lambda_1(1) = \mu_1(1)/\mu_1$ by using the relationship between the coefficients appearing in the various definitions of the rate function. Indeed, if $\ell(\vec{x}, \vec{y})$ is given by a particular $\vec{\theta}^*$ in (15.12), and by a particular set of $\lambda_1(1)$, $\lambda_2(1)$, and $\mu_1(1)$ in (15.14), then Exercise 5.29 gives

$$\lambda_1(1) = \lambda_1 e^{\langle \vec{\theta}^*, \vec{e}_1 \rangle}, \ \lambda_2(1) = \lambda_2 e^{\langle \vec{\theta}^*, \vec{e}_2 \rangle}, \ \mu_1(1) = \mu_1 e^{\langle \vec{\theta}^*, \vec{e}_3 \rangle}. \tag{15.17}$$

The relation now follows directly from Definition 15.2.

Relation (15.17) allows us to do partial calculations with each form of the rate function, and then to piece the calculations together. Let us define $C \stackrel{\triangle}{=} \lambda_1(1)/\lambda_1$. The quantity to be minimized on the right of (15.15) becomes, after some algebra,

$$\frac{(\lambda_1 + \lambda_2)(C \log C + 1 - C) + \mu_1 \left(\frac{1}{C} \log \frac{1}{C} + 1 - \frac{1}{C} \right)}{(\lambda_1 + \lambda_2)C - \mu_1/C}. \tag{15.18}$$

Define $\lambda \stackrel{\triangle}{=} \lambda_1 + \lambda_2$ and $\mu \stackrel{\triangle}{=} \mu_1$, and compare the minimization of (15.18) to the variational problem (11.8) of the $M/M/1$ queue, starting at zero and exiting at one, and its solution in Theorem 11.15 (see also Lemma C.8). Ignoring for a moment the constraint $\lambda_1(1) \ge \mu_1(1)$, the quantity to be minimized in expression (15.18) is the same one encountered in the $M/M/1$ queue, and we conclude that

the optimal change of measure is where the twisted arrival rate is μ and the twisted service rate is λ, or

$$C = \frac{\mu}{\lambda} = \frac{\mu_1}{\lambda_1 + \lambda_2}\,. \tag{15.19}$$

This corresponds to a path $\vec{r}^{\,*}(t)$ that travels in direction

$$\begin{aligned} \vec{r}^{\,*\prime}(t) &= \lambda_1 C\vec{e}_1 + \lambda_2 C\vec{e}_2 + \frac{\mu_1}{C}\vec{e}_3 \\ &= \left(\frac{\lambda_1\mu_1}{\lambda_1+\lambda_2} - (\lambda_1+\lambda_2), \frac{\lambda_2\mu_1}{\lambda_1+\lambda_2}\right), \end{aligned} \tag{15.20}$$

and by (11.23), the associated cost is $\log C$.

Perhaps the best way to interpret this equation is using the result of §7.4. There we showed that any path $\vec{r}(t)$ is equivalent to a change of measure $\lambda_i \to \lambda_i e^{\langle\vec{\theta},\vec{e}_i\rangle}$; in the present case,

$$\vec{\theta} = (\log C, \log C)$$

since $\lambda_i \to C\lambda_i$. The path defined in (15.20) clearly goes from $\vec{0}$ to E in time $T^* = (\mu_1 - \lambda_1 - \lambda_2)^{-1}$. The total cost of this path $I_0^{T^*}(\vec{r}^{\,*})$ is

$$I^* = I_0^{T^*}(\vec{r}^{\,*}) = \log\frac{\mu_1}{\lambda_1+\lambda_2}.$$

The only question is, is this new path truly in the interior of the state space? That is, under this twist, is $r_1^*(t) \geq 0$? From (15.20), $r_1^*(t) \geq 0$ for $t > 0$ if and only if

$$\lambda_2 < \lambda_c \stackrel{\triangle}{=} \sqrt{\mu_1\lambda_1} - \lambda_1. \tag{15.21}$$

Exercise 15.5. Derive (15.19) directly from the minimization (15.18), and verify the calculation leading to (15.21). ♠

We have now proved that the condition for the interior extremal path to be strictly away from the boundary is $\lambda_2 < \lambda_c$. In case of equality clearly it is better to move along the boundary. In Exercise 15.6 below we show that if $\lambda_2 > \lambda_c$ then it is also optimal to move along the boundary. Note that the condition does not depend on the value of μ_2; this is reasonable since μ_2 doesn't operate in the interior. Also note that λ_c has the right dimensions (the same as λ_1 and μ_1, inverse time).

Exercise 15.6. Show that if $\lambda_2 > \lambda_c$ then the optimal straight line path is along the boundary $x_1 = 0$. Hint: consider the last expression in (15.15) but add the constraint $\lambda_1(1) \geq \mu_1(1)$. Under the above condition there is no local minimum in the interior of the parameter set ($0 < \lambda_i(1)$, $\lambda_1(1) > \mu_1(1) > 0$). Use growth estimates on $h(a, b)$ near $a = 0$ and as $a \to \infty$ and continuity to conclude that there must be a minimum on the boundary $\lambda_1(1) = \mu_1(1)$. ♠

As before, we could obtain the first and third equalities as an immediate consequence of Exercise 15.8; but as we need all these, we shall do it the hard way.

Exercise 15.11. Verify (15.25). Hints: write π_0 as an explicit function of the parameters. Show that

$$(1-\pi_0)\frac{\partial \pi_0}{\partial \lambda_1(0)} = \pi_0 \frac{\partial \pi_0}{\partial \lambda_1(1)} = -\pi_0 \frac{\partial \pi_0}{\partial \mu_1(1)} \ .$$

Now equate derivatives to zero. You might also use (15.24). ♠

From the definition of y_2 in terms of the twisted rates we see that

$$y_2 = \rho\lambda_2 - \pi_0\mu_2/\rho. \tag{15.26}$$

We can now write the minimum cost in terms of ρ and π_0 alone and reduce the problem of finding the minimum cost path along the boundary $x_1 = 0$ to minimizing, over $0 \le \pi_0 \le 1$, $\rho > 0$, the function

$$h_b(\pi_0, \rho) \stackrel{\triangle}{=} \frac{\lambda_2 h(\rho, 1) + \pi_0\mu_2 h\left(\frac{1}{\rho}, 1\right) + U}{\rho\lambda_2 - \pi_0\mu_2/\rho}, \tag{15.27}$$

where we define

$$U \stackrel{\triangle}{=} \left(\sqrt{(1-\pi_0)\mu_1} - \sqrt{\lambda_1}\right)^2. \tag{15.28}$$

Note that U represents the cost of staying next to the boundary.

Exercise 15.12. Derive (15.27). Hint: use (15.25) and show that

$$(1-\pi_0)\big[h(\lambda_1, \lambda_1(1)) + h(\mu_1, \mu_1(1))\big] + \pi_0 h(\lambda_1, \lambda_1(0)) = U. \tag{15.29}$$

Now write C as a function of π_0. ♠

Note first that a minimal value of h_b can come only when $\pi_0 \in [0, 1-\lambda_1/\mu_1)$; this is because U is increasing for larger π_0 (note that $\sqrt{\mu_1(1-\pi_0)} - \sqrt{\lambda_1} = 0$ only when $\pi_0 = 1 - \lambda_1/\mu_1$), so that h_b is increasing in that range. The physical intuition is that there is no benefit to putting more effort into the balance in the x_1 direction, since the larger π_0 is, the harder it is for $x_2(t)$ to increase (the low-priority service rate is $\mu_2\pi_0$). For any given value of $\pi_0 \in [0, 1-\lambda_1/\mu_1)$, the value of ρ that minimizes h_b occurs where

$$\frac{\partial h_b}{\partial \rho} = 0 \text{ at } \rho = \frac{\lambda_2 + \pi_0\mu_2 + U + \sqrt{(\lambda_2 + \pi_0\mu_2 + U)^2 - 4\lambda_2\mu_2\pi_0}}{2\lambda_2}. \tag{15.30}$$

Exercise 15.13. Derive (15.30). Hint: show that $h_b \to \infty$ as $\rho \to \infty$ or as $\rho \downarrow \sqrt{\pi_0\mu_2/\lambda_2}$ (the point where $y_2 = 0$). Therefore there is a minimum of h_b where $\partial h_b/\partial\rho = 0$. Factor the resulting numerator into two quadratics, one of

which has no roots, so the roots of the other contains the correct value of ρ. Use $y_2 > 0$ in (15.26) to choose the correct root. You will find that

$$y_2 = \sqrt{(\lambda_2 + \pi_0\mu_2 + U)^2 - 4\lambda_2\mu_2\pi_0},$$

so that we can write

$$\rho = \frac{y_2 + \sqrt{y_2^2 + 4\lambda_2\pi_0\mu_2}}{2\lambda_2}.$$

You will recognize this $\rho = \exp(\theta^*)$, where θ^* is the optimal twist for an $M/M/1$ queue with arrival rate λ_2 and service rate $\pi_0\mu_2$ traveling with velocity y_2. ♠

By Exercise 15.13, the minimum value of h_b occurs, in the (ρ, π_0) space, for ρ bounded above and bounded away from zero. In order to see whether the minimum occurs at $\pi_0 = 0$ or at $\pi_0 > 0$, we look for a minimum of h_b, where $\rho(\pi_0)$ is given in (15.30). We have already argued that there could not be a minimum to the right of $1 - \lambda_1/\mu_1$, and since U and its derivative vanish there, this point cannot be a minimum. By continuity of h_b in π_0, there must be a minimum in $[0, 1 - \lambda_1/\mu_1)$. Define

$$\lambda_D \stackrel{\triangle}{=} \lambda_c \left(1 + \frac{\mu_2 - \mu_1}{\sqrt{\mu_1\lambda_1}}\right). \tag{15.31}$$

Lemma 15.14. *If $\lambda_2 > \lambda_D$ then the optimal value of h_b occurs at $\pi_0 > 0$. If $0 \le \lambda_2 \le \lambda_D$ then the optimal value of h_b occurs at $\pi_0 = 0$.*

Proof. The computations below establish the following.

(i) For each fixed ρ, the function $h_b(\pi_0, \rho)$ has only one local minimum for π_0 in the interval $[0, 1 - \lambda_1/\mu_1)$ that is hence the global minimum.

(ii) When $\rho(\pi_0)$ is given by (15.30), $\partial h_b/\partial\pi_0 = 0$ at $\pi_0 = 0$ only if $\lambda_2 = \lambda_D$.

(iii) If $\lambda_D \ge \lambda_2 = 0$ the minimum is at $\pi_0 = 0$, while if $\lambda_2 = \lambda_M \stackrel{\triangle}{=} \mu_2(1 - \lambda_1/\mu_1)$ then the minimum is at $\pi_0 > 0$.

By (ii)–(iii), if $\lambda_D > \lambda_2 = 0$ then $\partial h_b/\partial\pi_0 > 0$ at $\pi_0 = 0$, and by continuity in λ_2 and (ii) we conclude that the derivative is never negative for λ in $[0, \lambda_D]$. From (i) we now conclude that in this case the optimal value of π_0 is zero. Similarly, since by Exercise 15.17 $\lambda_M \ne \lambda_D$ we have from (i)–(iii) that $\partial h_b(0, \rho(0))/\partial\pi_0 < 0$ for $\lambda = \lambda_M$, and by continuity and (ii) it must stay negative for $\lambda > \lambda_D$. By (i) now we have that the optimal value of π_0 is strictly positive for $\lambda_2 > \lambda_D$. The result of the lemma now follows from (i) and the continuity in π_0, so that it remains to check (i)–(iii).

To establish (i), compute from (15.27) $\partial h_b/\partial\pi_0 = N/D$ by the usual rule for differentiating quotients. Since D is the square of the denominator y_2 in (15.27), we have $D > 0$. You might want to check the computations leading to (15.32); we used the computer algebra tool Theorist©, and will not write out the complicated

expressions. Repeating the procedure for N, denote $\partial N/\partial\pi_0 = C/D_2$. Again D_2 is positive, and

$$C = \frac{\sqrt{\mu_1\lambda_1}\left(\lambda_2\rho^2 - \mu_2\pi_0\right)\rho}{2\sqrt{1-\pi_0}}. \tag{15.32}$$

But we know that any value of ρ that might give us a minimum must have $y_2 > 0$ in (15.26). Therefore $C > 0$, and so $\partial N/\partial\pi_0 > 0$, so that N is an increasing function of π_0. This means that there is at most one root of $\partial h_b/\partial\pi_0$ in the interval $[0, 1-\lambda_1/\mu_1)$, and if $\partial h_b/\partial\pi_0 > 0$ at $\pi_0 = 0$, then $\pi_0 = 0$ must be the unique minimum. (Why? Think about it...) One of the two must hold, since the minimum cannot be at $1-\lambda_1/\mu_1$. This establishes (i).

To establish (ii), note from Exercise 15.13 that

$$\rho(0) = \frac{y_2}{\lambda_2} = \frac{\left(\sqrt{\mu_1} - \sqrt{\lambda_1}\right)^2}{\lambda_2} + 1. \tag{15.33}$$

(You can check that this gives $\rho = \sqrt{\mu_1/\lambda_1}$ when $\lambda_2 = \lambda_c$. This is the same as the twist for the interior path at $\lambda_2 = \lambda_c$; when $\pi_0 = 0$ the interior and boundary costs are hence the same, as they should be.) Now at this value of ρ we have $\partial h_b(0, \rho(0))/\partial\pi_0 = 0$ when $\lambda_2 = \lambda_D$.

Finally, in Exercise 15.18 below we show that $\pi_0 = 0$ is optimal for $\lambda_2 \approx 0$ provided $\lambda_D \geq 0$, and in Exercise 15.19 below we show that $\pi_0 > 0$ is optimal for $\lambda_2 \approx \lambda_M$, which implies (iii). ■

Exercise 15.15. Show that λ_D of (15.31) satisfies

$$\lambda_D = \frac{\left(\mu_2 - \mu_1 + \sqrt{\mu_1\lambda_1}\right)U}{\mu_1 - \sqrt{\mu_1\lambda_1}}.$$

(This may be the form in which you find λ_D, depending on how you do your calculations, in Exercise 15.16 below.) ♠

Exercise 15.16. Check that equations (15.31) and (15.33) are correct. ♠

Exercise 15.17. Prove that $\lambda_D < \mu_2\left(1 - \frac{\lambda_1}{\mu_1}\right) \stackrel{\triangle}{=} \lambda_M$, so that, for any given values of λ_1, μ_1, and μ_2, there always exists an interval of λ_2 where the cheapest path spends real time on the boundary. ♠

Exercise 15.18. If $\lambda_D \geq 0$ and $\lambda_2 \approx 0$ then the minimum of h_b is obtained at $\pi_0 = 0$. Hint: show the derivative in (15.27) at $\pi_0 = 0$ is positive. Note that when $\pi_0 = 0$, $\lambda_2 \cdot \rho \to U$ as $\lambda_2 \to 0$. ♠

Exercise 15.19. If $\lambda_2 \approx \lambda_M$ then the minimum of h_b is obtained at $\pi_0 > 0$. Hint: take $\pi_0 = 1-\lambda_1/\mu_1$ and $\lambda_2 = \mu_2\pi_0(1-x)$ with x small and positive. Use $\rho \approx 1 + x$ and $h(\rho, 1) \approx x^2/2$ to show $h_b(\pi_0, \rho) \approx 0$. ♠

of G. Suppose that for each $\vec{y} \in E$,

$$\mathbb{P}_{ss}(\vec{y}) = e^{-nV(\vec{y})+o(n)} \tag{15.52}$$

[recall $V(y) \stackrel{\triangle}{=} \inf\{I_0^T(\vec{r}) : \vec{r}(0) = \vec{q},\ \vec{r}(T) = \vec{y},\ T > 0\}$, where we assume that there exists a unique point $\vec{q} = \lim_t \vec{z}_\infty(t)$ for every starting point $\vec{z}_\infty(0)$]. Suppose further that

$$\mathbb{P}_{ss}(E) = e^{-n \inf_{\vec{x}\in E} V(\vec{x})+o(n)} \tag{15.53}$$

and suppose that there is a unique $\vec{y}^{\,*} \in E$ with

$$V(\vec{y}^{\,*}) = \inf_{\vec{x}\in E} V(\vec{x}). \tag{15.54}$$

Then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_{ss}\left(\left|\vec{z}_n(t) - \vec{y}^{\,*}\right| < \varepsilon \;\middle|\; \vec{z}_n(t) \in E \right) = 1. \tag{15.55}$$

Furthermore, if there exist constants n_0 and K_0 and a function $f(K) \to \infty$ as $K \to \infty$, such that

$$\mathbb{P}_{ss}\left(|\vec{z}_n(t)| \geq K\right) \leq e^{-nf(K)}, \tag{15.56}$$

then

$$\lim_{n\to\infty} \mathbb{E}_{ss}\left(\vec{z}_n(t) \;\middle|\; \vec{z}_n(t) \in E \right) = \vec{y}^{\,*}. \tag{15.57}$$

Exercise 15.42. Prove Theorem 15.39. ♠

The preemptive priority queue satisfies the hypotheses of Theorem 15.39 [even (15.56), since the process is bounded!]. This completes the justification of Theorems 15.31 and 15.32 for the preemptive priority queue and the set E.

Next we examine the justification of our results for the extended model, where there are two queues that are either bounded by n or are unbounded. For the unbounded queues we need exactly one more lemma: we need to see that (15.56) holds in order to use Theorem 15.39.

Lemma 15.41. *There exists a constant $C > 0$ and an n_0 such that for any $n > n_0$ and any $K \geq 1$,*

$$\mathbb{P}_{ss}\left(\{x_1 \geq nK\} \bigcup \{x_2 \geq nK\} \right) \leq e^{-nKC}. \tag{15.58}$$

Exercise 15.42. Prove Lemma 15.41. Hint: Kurtz's Theorem applies to the process. ♠

These paths have costs $I_0^T(\vec{r}^{\,*})$ given by $2\log(\mu/2\lambda)$ and $\log(\mu/\lambda)$ respectively. Therefore, for $4\lambda > \mu$, the first path is cheaper (more likely), whereas for $4\lambda < \mu$ the second path is cheaper. When $4\lambda = \mu$ both paths have the same cost.

Now let's make sure that we have considered all possible critical paths. Since the process is constant coefficient in the regions $x_1 > x_2$ and $x_1 < x_2$, we know that the critical paths must consist of straight-line segments. All possibilities are sketched in Figure 15.56.

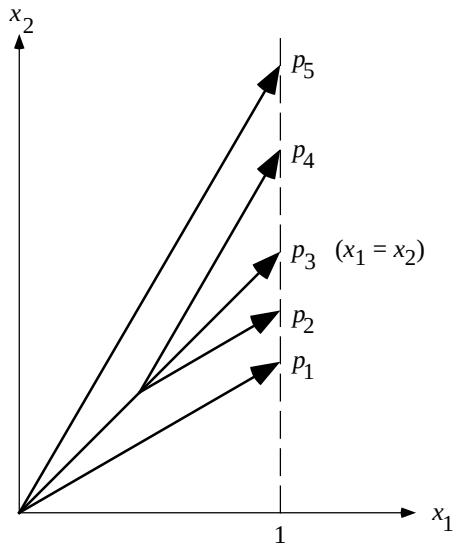

Figure 15.56. The possible critical paths.

The point p_1 represents the path $\vec{r}_1^{\,*}$ that goes from $(0, 0)$ to $x_1 = 1$ with cost $\log(\mu/\lambda)$ in the interior defined in (15.82). The point p_3 represents the minimum cost path $\vec{r}_b^{\,*}$ that travels along the boundary, defined in (15.79). Its cost is given in (15.80) as $2\log(\mu/2\lambda)$. The point p_5 represents the path you calculated in Exercise 15.55, where $x_2(t) > x_1(t)$.

Since the paths are straight lines and the process is constant-coefficient in each region, the cost incurred per unit time along each critical path is $\ell(r, r')$, a constant. Therefore there is a constant cost per unit distance traveled toward $x_1 = 1$. Now if the costs of p_1 and p_3 are the same (that is, if $4\lambda = \mu$), then any path (such as $\vec{0}$—p_2 in Figure 15.56) that follows p_3 for a while and then travels parallel to p_1 thereafter will have the same cost. Similar considerations apply to constructing p_4 from p_3 and p_5, although p_4 and p_5 are never minimal paths, by Exercise 15.55. Again, the case where p_2 is a minimal path is exactly the critical case $4\lambda = \mu$ derived above. This finishes the proof of Theorem 15.45.

Leo Flatto [Fl] has derived the steady state of this system; indeed, his work motivated this study. He found that when $4\lambda = \mu$, the distribution of x_2 conditioned on $x_1 = 1$ is asymptotically (as $n \to \infty$) *uniform* on $x_2 \in \left(\frac{\lambda}{\mu - \lambda}, 1\right)$. Our interpre-

Transformation and most probable path.

Although $\vec{z}_n(t)$ does not fall under the domain of flat boundary theory, just as in §15.9 a simple transformation of the state space will enable us to calculate (and prove) limit laws for $\vec{z}_\infty(t)$ and the large deviations principle for $\vec{z}_n(t)$. We use the transformations R and $\mathbf{R}$ defined in (15.66), §15.9, and let $\vec{u} = \mathbf{R}\vec{x}$. In the notation of Chapter 8, u_2 was denoted u_0 and is the direction perpendicular to the boundary.

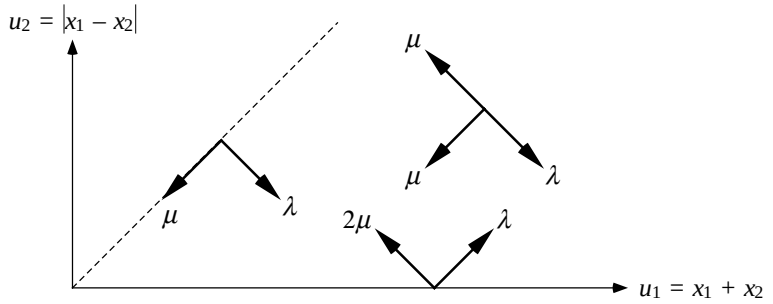

Figure 15.59. The transformed state space.

Then it is a simple matter to calculate the equation of $\vec{u}_\infty$, the fluid limit for the process $\vec{u}_n(t)$ (as in §15.9, $\vec{u}_n$ is the scaled process derived from $\vec{u}$):

$$\frac{d}{dt}\vec{u}_\infty(t) = \begin{cases} (-\mu, -\mu) & \text{if } u_1 = u_2 > 0 \text{ and } \lambda < \mu \\ (\lambda - 2\mu, -\lambda) & \text{if } u_1 = u_2 > 0 \text{ and } \lambda \geq \mu \\ (\lambda - 2\mu, -\lambda) & \text{if } u_1 > u_2 > 0 \\ (\lambda - 2\mu, 0) & \text{if } u_1 > 0 \text{ and } u_2 = 0 \\ \vec{0} & \text{if } \vec{u} = \vec{0}. \end{cases} \tag{15.83}$$

Here is how we derive (15.83): the computations are left to Exercise 15.60 below. In the interior of the state space, where $u_1 > u_2 > 0$, the third line of (15.83) is just (5.7), which clearly follows from Theorem 5.3. Depending on $\vec{u}_\infty(0)$ and whether or not $\lambda > \mu$, this path may lead to the line $u_1 = 0$ or to the line $u_1 = u_2$ (or it may lead directly to $\vec{0}$, which we analyze below). The line $u_1 = u_2$ will be reached by $\vec{u}_\infty$ only if $\lambda < \mu$. Once $\vec{u}_\infty$ hits either of the boundaries, we calculate its subsequent evolution by Exercise 8.13. This leads to the first and fourth lines of (15.83). Looking at the solutions of (15.83) shows that $\vec{u}_\infty \to \vec{0}$ if and only if $\lambda < 2\mu$. We could have guessed this stability condition—the total number of customers in the system is pretty much the same as an $M/M/1$ queue with arrival rate λ and service rate 2μ. The only difference is when one of the queues is empty: then the rate at which the system empties may only be μ.

Exercise 15.60. Prove that (15.83) is the most likely behavior away from $\vec{0}$. Hint: this is a local result, so the theory of Chapter 8 applies away from $\vec{0}$. ♠

Exercise 15.61. Derive $\vec{z}_\infty(t)$ from $\vec{u}_\infty(t)$. ♠

1. Upcrossings are rare. (We know this since the total queue length is like a stable $M/M/1$ queue. We can also provide a supermartingale that shows positive recurrence; this, combined with the next item, means that upcrossings must be rare.)
2. The process has bounded speed (the λ_i are bounded, not necessarily continuous).
3. We know where upcrossings occur most likely.
4. The process is ergodic (positive recurrent). This is straightforward. As mentioned in point 1, there is a simple supermartingale that shows this: the process $r(t)$, the Euclidean distance between $\vec{0}$ and $\vec{x}(t)$, is a supermartingale if the stability condition holds and if $|\vec{x}|$ is not too small (this is an easy calculation). ■

Exercise 15.63. Show that, for the problem of finding when $x_1 \geq n$, the optimal path is

$$\vec{r}^{\,*}(t) = \left(\mu - \frac{\lambda}{2}, \mu - \frac{\lambda}{2}\right) t \quad \text{for} \quad 0 \leq t \leq T^* = \frac{1}{\mu - \lambda/2}.$$

[This is pretty obvious from the transition diagram for $\vec{u}(t)$: we see that $u_1(t)$ looks like an $M/M/1$ queue with arrival rate λ and service rate 2μ, so that most likely upcrossings are made by the arrival and service rates switching:

$$u_1(t) = (x_1 + x_2)(t) = (2\mu - \lambda) \cdot t.$$

Also, during this excursion,

$$u_2(t) = |x_1 - x_2| \approx 0.$$

This gives the stated $\vec{r}^{\,*}(t)$.] The exercise is to write the cost function for $\vec{u}(t)$, and solve the appropriate variational problem. ♠

Extensions:

1. The queues may have different service rates. Nothing changes so long as we have $\lambda < \mu_1 + \mu_2$. That is, the conclusion of Theorem 15.62 holds.
2. There can be any number of queues. Each queue has its own service rate μ_i. Customers always join the shortest queue, choosing uniformly at random among all shortest queues at the time of their arrival. We assume the stability condition $\lambda < \sum_i \mu_i$.

Proof 1 breaks down for these extensions, but proof 2 extends verbatim. The reasons that proof 1 breaks down are manifold. We don't have a large deviations principle for processes with discontinuities along the boundary such as an edge. Furthermore, the mapping from $\vec{x}$ to $\vec{u}$ no longer gives a nice mapping of the rates. That is, consider the jump rates for $\vec{u}$ when $\mu_1 \neq \mu_2$ and $x_1 > x_2$; then consider the jump rates when $x_1 < x_2$. Since this mapping is no longer one-to-one, we cannot use a large deviations principle for $\vec{u}_n$ to establish one for $\vec{z}_n$.

Exercise A.30. Establish the following. If f is lower semicontinuous than $(-f)$ is upper semicontinuous. A function that is both lower semicontinuous and upper semicontinuous is continuous. The supremum of a family of lower semicontinuous functions is lower semicontinuous. The composition $f(g(x))$ of continuous functions is continuous. Hint: use the definitions, not the theorems. ♠

Theorem A.31. *A lower semicontinuous function attains its minimum on every compact set.*

Definition A.32. *A function of a real variable is right continuous if*

$$\lim_{x \downarrow y} f(x) = f(y),$$

that is, $f(x)$ converges to $f(y)$ as x decreases to y. The definition of left continuous is analogous. (This is generalized in Definition A.52.)

Definition A.33. *A set G in $\mathbb{R}^d$ has smooth boundary if there is a function f from $\mathbb{R}^d$ to $\mathbb{R}$ that is twice continuously differentiable, so that $f(\vec{x}) = 0$ implies $|\nabla f(\vec{x})| \neq 0$, and*

$$\vec{x} \in \partial G \quad \text{if and only if} \quad f(\vec{x}) = 0.$$

Lemma A.34. *Let f be a real-valued function of the variables $x \in A$ and $y \in B$, and $y_1, y_2, \ldots$ an arbitrary sequence in B. Then*

(i)
$$\begin{aligned} \liminf_n \inf_{x \in A} f(x, y_n) &\leq \inf_{x \in A} \liminf_n f(x, y_n) \\ \limsup_n \inf_{x \in A} f(x, y_n) &\leq \inf_{x \in A} \limsup_n f(x, y_n) \\ \sup_{y \in B} \inf_{x \in A} f(x, y) &\leq \inf_{x \in A} \sup_{y \in B} f(x, y). \end{aligned}$$

(ii) If, in addition, the following two equalities hold:

$$\inf_{x \in A} \sup_{y \in B} f(x, y) = \sup_{y \in B} \inf_{x \in A} f(x, y) = \lim_n \inf_{x \in A} f(x, y_n),$$

then $$\lim_n \inf_{x \in A} f(x, y_n) = \inf_{x \in A} \liminf_n f(x, y_n) = \inf_{x \in A} \limsup_n f(x, y_n).$$

Proof. Denote $f_i(y) \stackrel{\triangle}{=} \inf_{x \in A} f(x, y)$. Then $f_i(y) \leq f(x, y)$ for all x, and so

$$\liminf_n f_i(y_n) \leq \liminf_n f(x, y)$$

for all x, and hence

$$\liminf_n f_i(y_n) \leq \inf_{x \in A} \liminf_n f(x, y),$$

which is exactly (i). The proof with lim sup or sup is identical.

From (i) we obtain

$$\liminf_n \inf_{x \in A} f(x, y_n) \leq \inf_{x \in A} \liminf_n f(x, y_n) \leq \inf_{x \in A} \sup_{y \in B} f(x, y)$$

that if $A_1, A_2, \ldots$ are disjoint sets in $\mathcal{F}$ then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu\left(A_i\right).$$

The triple $(\Omega, \mathcal{F}, \mu)$ is then called a measure space. A measure μ is σ-finite if there are sets $A_1, A_2, \ldots$ in $\mathcal{F}$ so that

$$\bigcup_{i=1}^{\infty} A_i = \Omega \quad \textit{and} \quad \mu(A_i) < \infty.$$

If in addition $\mu(\Omega) = 1$ then μ is called a probability measure, and $(\Omega, \mathcal{F}, \mu)$ is called a probability space. For a set A in $\mathcal{F}$, we say A holds μ-a.s. ("almost surely") if $\mu(A^c) = 0$. If μ is a probability measure then the equivalent term "w.p. 1" (with probability one) is used.

Throughout the book we consider only σ-finite measures.

Exercise A.71. Let A_n be events on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Show that for all a and b, $\lim_{n\to\infty} \frac{1}{n} \log\left(e^{na} + e^{nb}\right) = \max\{a, b\}$, and for all positive a and ε,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(A_n\right) \leq -a \quad \text{implies} \quad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(A_n^c\right) \geq -\varepsilon. \qquad \spadesuit$$

Definition A.72. *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces. The product space $\Omega_1 \times \Omega_2$ is the collection of all pairs $\left(\{\omega_1, \omega_2\} : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\right)$. The product σ-field $\mathcal{F}$ is the smallest σ-field that includes all sets of the form $\left\{\{\omega_1, \omega_2\} : \omega_1 \in A_1,\ \omega_2 \in A_2\right\}$ for all sets $A_i \in \mathcal{F}_i$. The product measure μ is the (unique) measure on the product space for which $\mu(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2)$ for all sets $A_i \in \mathcal{F}_i$.*

Definition A.73. *Let $(\mathcal{X}, d)$ be a Polish space. The smallest σ-field of subsets of $\mathcal{X}$ that contains all open balls is called the Borel σ-field and is denoted $\mathcal{B}(\mathcal{X})$. Sets in $\mathcal{B}(\mathcal{X})$ are called Borel sets.*

Example A.74. There exists a unique measure m on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the property that $m([a, b]) = b - a$ for all $b > a$. It is called Lebesgue measure. A set A has a Lebesgue measure zero if and only if for any $\varepsilon > 0$ it can be covered by a countable number of intervals, the sum of whose lengths is less than ε.

In fact, the Lebesgue measure is defined on a collection $\mathcal{F}$ of Lebesgue measurable sets that (strictly) contains $\mathcal{B}(\mathbb{R})$. We shall not discuss the resulting fine points: see any book on real analysis or modern integration theory.

Definition A.75. *We say a property holds almost everywhere (abbreviated as a.e.) if it holds outside a set of Lebesgue measure zero. A function $f : \mathbb{R} \to \mathbb{R}^d$ is called essentially bounded if there exists a bound B so that $|f| \leq B$ a.e.*

Throughout the book we hold a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ fixed, where $\mathbb{P}$ is a probability measure. But note that the measures we consider below are not necessarily probability measures, and in particular we might have $\mu(\Omega) = \infty$.

$$T_t \geq 0 \tag{B.4}$$
$$T_t 1 = 1 \tag{B.5}$$
$$T_0 = I \tag{B.6}$$
$$\lim_{t \searrow 0} T_t = I \tag{B.7}$$

[where the first condition means that $\sum_y p_t(x, y) p_s(y, z) = p_{t+s}(x, z)$]. Equation (B.4) means the operator T_t maps positive functions to positive functions. Intuitively, the quantity $p_t(x, y)$ represents the probability that at time t the process will be at y given that at time zero it was at x.

There exists a one-to-one correspondence between generators and transition semigroups. Namely, given a generator L, the corresponding semigroup is given by

$$T_t = e^{tL} \stackrel{\triangle}{=} I + tL + \ldots + \frac{(tL)^n}{n!} + \ldots$$

Conversely, given a transition semigroup T_t, the generator is

$$L = \lim_{t \searrow 0} \frac{T_t - I}{t} \tag{B.8}$$

Uniformization.

There is one more representation for discrete-space continuous-time Markov processes that works when the jump rates are uniformly bounded: $\max_x |a(x, x)| < \infty$. Let λ be a positive real number that dominates these rates and put

$$Q = \lambda^{-1} L + I.$$

It is easy to see that Q is nonnegative and has row sums equal to one (i.e., it is a stochastic matrix). A simple calculation shows that

$$T_t = e^{tL} = e^{-\lambda t} e^{\lambda t Q} = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} Q^k. \tag{B.9}$$

This is called the uniformization theorem. The interpretation of the right-hand side is that there is a single exponential clock that ticks at rate λ and every time it ticks, a jump occurs according to the stochastic matrix Q. Hence, at time t, the number of times this clock has ticked is a Poisson random variable with parameter λt and so the chance that there have been k jumps is the Poisson probability that appears in front of Q^k in (B.9). If this result seems surprising (in particular the fact that λ can be chosen arbitrarily large), it might help to bear in mind that the stochastic matrix Q allows jumps from a state x to itself.

$$
\begin{aligned}
&= r'(t)\frac{\partial}{\partial r}\, f(r(t), r'(t)) + r''(t)\frac{\partial}{\partial r'}\, f(r(t), r'(t)) \\
&\qquad - r''(t)\frac{\partial}{\partial r'}\, f(r(t), r'(t)) - r'(t)\frac{d}{dt}\frac{\partial}{\partial r'}\, f(r(t), r'(t)) \\
&= r'(t)\left[\frac{\partial}{\partial r}\, f(r(t), r'(t)) - \frac{d}{dt}\frac{\partial}{\partial r'}\, f(r(t), r'(t))\right] \\
&= 0.
\end{aligned}
$$

Therefore, (C.2) implies the DuBois-Reymond Equation:

$$
f(r(t), r'(t)) - r'(t)\frac{\partial}{\partial r'}\, f(r(t), r'(t)) = \text{constant.} \qquad (C.3)
$$

Next consider the case where b is free, that is, we wish to minimize I over all r with $r(a) = \alpha$ and $r(b) = \beta$, but also over all possible values of b (assume $\alpha \neq \beta$). Then the Euler Equation (C.2) is still a necessary condition, obtained from the optimality property in r at the best value of b. We now derive an additional equation, obtained from the optimality in b around the best r. We need to be careful here that we only consider functions satisfying the boundary conditions! So assume r is defined on $[a, b+1]$ and consider $\delta > 0$; the calculation for $\delta < 0$ is almost identical. By definition $r(b) = \beta$: to satisfy the boundary condition at $b + \delta$ we need to to choose the perturbation function so as to compensate for the fact that, generally, $r(b + \delta) \neq \beta$. So let δx be a function with

$$
\delta x(0) = 0, \quad \delta x(b + \delta) = -\delta \cdot r'(b).
$$

Then $r(t) + \delta x(t)$ satisfies the boundary conditions

$$
r(b + \delta) + \delta x(b + \delta) \approx r(b) + \delta \cdot r'(b) + \delta x(b + \delta) = \beta
$$

(at least to first-order approximation), and δx can (and will) be chosen linear in δ. [Note that we cannot simply choose $r(t) = r(b),\ t > b$ as this would not work for $\delta < 0$.] Taking derivatives and equating to zero now yields

$$
\begin{aligned}
&\frac{d}{d\delta}\int_a^{b+\delta} f\big(r(t) + \delta x(t), r'(t) + \delta x'(t)\big)\, dt \,\Big|_{\delta=0} \\
&\qquad = f(r(b), r'(b)) + \int_a^b \frac{d}{d\delta}\, f\big(r(t) + \delta x(t), r'(t) + \delta x'(t)\big)\, dt \,\Big|_{\delta=0}.
\end{aligned}
$$

But by the argument in the case of fixed b we know that the second term is zero for all x, except that in the integration by parts we no longer have $\delta x(b + \delta) = 0$. Thus we obtain

$$
\begin{aligned}
&\frac{d}{d\delta}\int_a^{b+\delta} f\big(r(t) + \delta x(t), r'(t) + \delta x'(t)\big)\, dt \,\Big|_{\delta=0} \\
&\qquad = f(r(b), r'(b)) + x(b) \cdot \frac{\partial}{\partial r'}\, f(r(b), r'(b))
\end{aligned}
$$

and substituting for the value of $x(b)$ we obtain

$$f(r(b), r'(b)) - r'(b) \cdot \frac{\partial}{\partial r'} f(r(b), r'(b)) = 0. \tag{C.4}$$

This condition is called the "transversality condition."

C.2. Calculus of Variations and Large Deviations

The reason we want to solve problems in the calculus of variations is that the solutions will answer questions about probabilities. However, solving problems in the calculus of variations is notoriously difficult. There are questions of existence of solutions, and most of the results are characterizations (necessary conditions), as will become clear in §C.4. But don't despair! We are fortunate on several counts. First, some properties of the rate function that we derived will address the existence problem, and many special tricks will come to the rescue.

Our approach to the variational problems will go as follows. First, the questions we are interested in are such that existence will be automatic due to lower-semicontinuity and compactness (see Theorem C.1 below). Then, we will usually derive a solution to the Euler equation. Using a variety of techniques, some from the calculus of variations and some specific to the problem under consideration, we will then show that this particular solution of the Euler equation is the only one that can be optimal. This will nail the validity of our formal calculation. Here is an obvious, if somewhat abstract theorem that formalizes this argument.

Theorem C.1. *Assume I is a good rate function, and consider the variational problem*

$$I^*(F) = \inf\{I(r) : r \in F\},$$

where F is a closed set so that $I^*(F) < \infty$. Suppose there is a set S so that

$$\mathcal{O} \stackrel{\triangle}{=} \left\{r : r \in F,\ I(r) = I^*(F)\right\} \subset S.$$

Then $\mathcal{O}$ is not empty, and if $S \bigcap F$ contains exactly one point r^* then r^* solves the variational problem. That is, $r^* \in F$ and $I(r^*) = I^*(F)$.

Remark. The set S defines "necessary conditions" for an optimum, in the sense that any optimal point must be in S. However, S may contain many nonoptimal points, and even points outside F.

Proof. By Exercise 2.3, there exists (at least one) point r^* in $\mathcal{O}$. By the hypothesis, every such point must be in $S \bigcap F$. But there is only one such point r^*. ■

The next two results concern the function ℓ, defined in Equations (5.2)–(5.4) [or in Theorem 8.19(i), with the obvious definition for g]. They will complement the Euler equations in making the set S of Theorem C.1 small enough so that the theorem can be applied.

Theorem C.2. *Assume $\ell(x, y)$ is differentiable in y and that*

$$\ell(x, y) = \theta^* y - g(x, \theta^*). \tag{C.5}$$

If the optimal point $\theta^* = \theta^*(x, y)$ is differentiable in y and $g(x, \theta)$ is differentiable in θ near θ^*, then

$$\frac{d}{dy}\ell(x, y) = \theta^*(x, y). \tag{C.6}$$

Proof. Taking derivatives we obtain

$$\begin{aligned}\frac{d}{dy}\ell(x, y) &= \frac{d}{dy}\left(\theta^* y - g(x, \theta^*)\right) \\ &= \theta^* + \left[y - \frac{\partial}{\partial\theta} g(x, \theta^*)\right]\frac{\partial}{\partial y}\theta^*(x, y) \\ &= \theta^*\end{aligned}$$

since the middle term in square brackets is zero as θ^* is a maximizer. ■

Theorem C.3. *Assume $\ell(x, y)$ and $\theta^*(x, y)$ satisfy (C.5) and (C.6). Then (C.3) implies [for $(x, y) = (r(t), r'(t))$] that $g(x, \theta^*) = C$, for some constant C. If in addition (C.4) holds then $C = 0$.*

Proof. By (C.3), for some constant C

$$y\frac{d}{dy}\ell(x, y) = \ell(x, y) + C = \theta^* y - g(x, \theta^*) + C$$

by definition of θ^*. But by (C.6) of Theorem C.2, the left-hand side equals $\theta^* y$ and the first result follows. If (C.4) holds then by the same argument $C = 0$. ■

In Chapters 5 and 8 we established that the rate functions I_0^T we deal with are indeed good rate functions. But our applications often involve "free time" problems. To be able to use the reasoning of Theorem C.1, we need to establish a priori that it suffices to look at finite times. So, consider a set G of pairs $(\vec{r}, T_r)$ so that each function $\vec{r}$ is defined on the interval $[0, T_r]$. With some abuse of notation, we will write $r \in G$ when we mean $(r, T_r) \in G$. Let F_G be the following set of functions on $[0, \infty]$. A function $\vec{s}$ is in F if and only if $\vec{s}(t) = \vec{r}(t),\ 0 \le t \le T_r$ for some $\vec{r}$ in G, and in addition $\vec{s}$ follows $\vec{z}_\infty$ for $T_r \le t < \infty$. Thus functions in F_G are just extensions of functions in G. Finally, let F_G^T be the collection of functions in F_G but with each function restricted to the interval $[0, T]$.

Lemma C.4. *Suppose that I_0^T is a good rate function for each $T < \infty$. Assume that there exists a T' so that the following holds: for each $\vec{r} \in G$ there is a $\tilde{r} \in G$ so that $T_{\tilde{r}} \le T'$ and $I_0^{T_{\tilde{r}}}(\tilde{r}) \le I_0^{T_r}(\vec{r})$. Then*

$$\begin{aligned}\inf\left\{I_0^{T'}(\vec{r}) : \vec{r} \in F_G^{T'}\right\} &= \inf\left\{I_0^{T_r}(\vec{r}) : \vec{r} \in G,\ T_r \le T'\right\} \\ &= \inf\left\{I_0^{T_r}(\vec{r}) : \vec{r} \in G\right\}.\end{aligned}$$

If $F_G^{T'}$ is closed then the infimum is attained.

Proof. If $\vec{s} \in F_G$ is an extension of $\vec{r} \in G$ and $T_r \leq T'$ then $I_0^{T'}(\vec{s}) = I_0^{T_r}(\vec{r})$, since ℓ vanishes along z_∞ paths. The first claim now follows from the assumptions. The last claim follows since by assumption, $I_0^{T'}$ is a good rate function. ■

C.3. One-Dimensional Tricks

Most of our variational problems turn out to be one-dimensional. In that case, several tricks are available which simplify the computations considerably—often to the point where analytic solutions are possible. We start with the more general properties. Consider the following level crossing problem. Given α, β,

$$\begin{aligned} &\text{minimize} \quad I_0^T(r) \triangleq \int_0^T f\big(r(t), r'(t)\big)\, dt \\ &\text{subject to} \quad T > 0,\ r(0) = \alpha,\ r(T) = \beta,\ r \text{ absolutely continuous,} \end{aligned} \tag{C.7}$$

where $T = \infty$ is acceptable. We denote the minimal value by I^* or $I^*(\alpha, \beta)$, and call this the value of the problem. We say that r^* solves the variational problem (C.7) if r^* satisfies the boundary conditions at zero and T, and $I_0^T(r^*) = I^*$.

Lemma C.5. *Consider the problem (C.7), and assume that $f(x, y) \geq 0$ for each x, y. If $\alpha < \beta$ then it suffices to consider strictly increasing functions. If $\beta < \alpha$ then it suffices to consider strictly decreasing functions.*

Proof. Pick any absolutely continuous function satisfying $r(0) = \alpha$ and $r(T) = \beta$, and assume that $r(u) < r(s)$ for some $0 \leq s < u \leq T$. Since r is continuous, there is some $u' \in [u, T)$ so that $r(u') = r(s)$. Since $f(x, y) \geq 0$ for each x, y, the function

$$\tilde{r}(t) \triangleq \begin{cases} r(t) & t \leq s, \\ r(t + u' - s) & t \geq s, \end{cases}$$

defined on the interval $[0, S]$ where $S = T - (u' - s)$ satisfies the boundary conditions and $I_0^S(\tilde{r}) \leq I_0^T(r)$. Thus we can modify any given function r until it is strictly increasing, without increasing the corresponding integral. Note that by absolute continuity, the number of required modifications is at most countable, so that the change in the derivative of r at these points will not affect the integral. Now repeat the same construction on intervals where r is constant. The proof when $\beta < \alpha$ is identical. ■

For a strictly increasing absolutely continuous function v, denote

$$t(x) = \{t : v(t) = x\} \quad \text{and} \quad v_t'(x) \triangleq \left[\frac{dv(t)}{dt}\right]_{t=t(x)}. \tag{C.8}$$

Note that this is well defined since v is strictly increasing.

is well defined for all $\alpha \le r \le \beta$ as y is bounded below, and defines a function $t(r)$ that is strictly monotone increasing. But then the inverse function $v(t)$ exists, and is the strictly monotone function determined by $t[v(t)] = t$. Both functions are differentiable due to the bounds on y, and differentiating

$$1 = \frac{dt}{dt} = \frac{dt(v(t))}{dr}\frac{dv(t)}{dt} = \frac{1}{y(v(t))}\frac{dv(t)}{dt}$$

so that v satisfies the desired relation. Finally, the defining relation gives $v(0) = \alpha$, and since y is bounded away from zero, there is some T so that $v(T) = \beta$. ■

In the parlance of convex analysis, if ℓ is not differentiable at $y(x)$ then zero belongs to the set of values that are obtained by replacing the derivative with a subgradient. Such points are isolated and do not affect the analysis. In our applications the functions will be differentiable, so we do not elaborate on this point.

Under the conditions of Lemma C.8 the conditions of Lemma C.6 are satisfied, so that v solves the variational problem (C.7). We are now ready to derive an expression for the value of the variational problem, as well as a method to compute an optimal solution.

Lemma C.9. *Consider problem (C.7) with $\alpha < \beta$ and where ℓ is given through (5.2)–(5.4). Assume that all points $x \in (\alpha, \beta)$ are nondegenerate, and that ℓ is differentiable in y at all points x and $y > 0$. For each x, the equation*

$$g(x, \theta^*) = 0 \tag{C.11}$$

has at most two solutions. If in addition all the conditions of Lemma C.8 hold, then with the appropriate choice of $\theta^*(x)$,

$$I^*(\alpha, \beta) = \int_\alpha^\beta \theta^*(x)\, dx. \tag{C.12}$$

Proof. Since $g(x, \theta)$ is strictly convex in θ, (C.11) has at most two solutions. Let $\theta^*(x, y)$ be a value of θ that achieves the maximum in the definition (5.2)–(5.4) of ℓ: that such a value exists (and is finite) follows from Lemma C.7. From that lemma it also follows that (C.6) holds, that is

$$\frac{\partial}{\partial y}\ell(x, y) = \theta^*(x, y) \tag{C.13}$$

for all x, y. This and the definition of ℓ yield

$$\ell(x, y) - y\frac{\partial \ell(x, y)}{\partial y} = -g(x, \theta^*(x, y)) \tag{C.14}$$

for all x, y. Now under the conditions of Lemma C.8 there exists an extremizing path $v(t)$, and from (C.10), along this path (that is, for all x, y so that, for some t, $v(t) = x$ while $v'(t) = y$),

$$\ell(x, y) - y\frac{\partial \ell(x, y)}{\partial y} = 0 \tag{C.15}$$

Theorem D.11 Gärtner-Ellis**.** *Let $\vec{s}_1, \vec{s}_2, \ldots$ be random variables with values in $\mathbb{R}^d$ so that Assumption D.3 holds. Then*

(i) For every closed set F

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{\vec{s}_n}{n} \in F\right) \leq -\inf_{\vec{x}\in F} I(\vec{x}).$$

(ii) For every open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{\vec{s}_n}{n} \in G\right) \geq -\inf_{\vec{x}\in G\cap\mathcal{F}} I(\vec{x}).$$

(iii) Assume that ϕ is differentiable in $\mathcal{D}_\phi^o$, and that $\mathcal{D}_\phi^o$ is non-empty. Assume further that ϕ is steep, that is if $\vec{\theta}$ is on the boundary of $\mathcal{D}_\phi$, then

$$\vec{\theta}_n \to \vec{\theta} \quad \textit{implies} \quad |\nabla\phi(\vec{\theta}_n)| \to \infty.$$

If ϕ is also lower semicontinuous, then the large deviations principle holds with a good rate function I.

Note that, unless strong assumptions are made on ϕ, it is possible that the bounds are not tight:

Exercise D.12. Show that if s_n is an exponential random variable with mean n, then the lower bound in Theorem D.11 is not tight. Hint: the set $\mathcal{F}$ contains only the point 0. ♠

As an immediate application, one can obtain a large deviations principle for finite-state Markov chains.

Example D.13. Let $y_1, y_2, \ldots$ be an ergodic Markov chain with a finite state space, say $1, 2, \ldots N$, and with transition matrix $\Pi = \{p_{ij}\}$. Let $\vec{x}_1, \vec{x}_2, \ldots$ be defined through $\vec{x}_n = f(y_n)$ for some function f, with values in $\mathbb{R}^d$. Define

$$p_{ij}^{\vec{\theta}} \stackrel{\triangle}{=} p_{ij} e^{\langle\vec{\theta}, f(j)\rangle}.$$

Let $\Pi_{\vec{\theta}} \stackrel{\triangle}{=} \left\{p_{ij}^{\vec{\theta}}\right\}_{i,j=1}^N$ and let $\rho(\Pi_{\vec{\theta}})$ denote the largest eigenvalue of the matrix $\Pi_{\vec{\theta}}$. Define

$$I(\vec{x}) \stackrel{\triangle}{=} \sup_{\vec{\theta}} \left(\langle\vec{\theta}, \vec{x}\rangle - \log\rho(\Pi_{\vec{\theta}})\right).$$

Then the sample mean $\vec{z}_n = n^{-1}\sum_{k=1}^n \vec{x}_k$ satisfies a large deviations principle with good, convex rate function I. For a proof of this statement, see [DZ, §3.1]. This can be shown [DZ, §3.2] to be equivalent to the level II result we obtain (using a different method) in §D.4.